Action Prediction During Human-Object Interaction Based on Dtw and Early Fusion of Human and Object Representations

Victoria Manousaki, Konstantinos Papoutsakis and Antonis Argyros

August 1, 2021

ACTION PREDICTION DURING HUMAN-OBJECT INTERACTION BASED ON DTW AND EARLY FUSION OF HUMAN AND OBJECT REPRESENTATIONS *

Victoria Manousaki[1,2], Konstantinos Papoutsakis[2], and Antonis Argyros[1,2]

[1] Computer Science Department, University of Crete
[2] Institute of Computer Science, Foundation for Research and Technology - Hellas (FORTH)
{vmanous,papoutsa,argyros}@ics.forth.gr

Abstract. Action prediction is defined as the inference of an action label while the action is still ongoing. Such a capability is extremely useful for early response and further action planning. In this paper, we consider the problem of action prediction in scenarios involving humans interacting with objects. We formulate an approach that builds time series representations of the performance of the humans and the objects. Such a representation of an ongoing action is then compared to prototype actions. This is achieved by a Dynamic Time Warping (DTW)-based time series alignment framework which identifies the best match between the ongoing action and the prototype ones. Our approach is evaluated quantitatively on three standard benchmark datasets. Our experimental results reveal the importance of the fusion of human- and object-centered action representations in the accuracy of action prediction. Moreover, we demonstrate that the proposed approach achieves significantly higher action prediction accuracy compared to competitive methods.

Keywords: Action Prediction · Early Action Classification · Dynamic Time Warping (DTW) · Soft DTW · Global Alignment Kernel (GAK).

1 Introduction

The capability to foresee and predict future outcomes is a very important human ability as it enables us to plan our next moves and actions and articulate with the environment. Likewise, prediction is an important capability of several technical systems. For example, home assisting robots need to be able to predict user actions and intentions in order to plan ahead their own actions and provide fast and accurate assistance to the user [9, 10]. In another example, autonomous cars [23] need the ability to predict the movement of pedestrians and other vehicles to plan a safe route and avoid accidents.

* The research work was supported by the Hellenic Foundation for Research and Innovation (HFRI) under the HFRI PhD Fellowship grant (Fellowship Number: 1592.)

In this paper, we investigate the specific problem of vision-based action prediction in human-object interaction scenarios. Action prediction is the capability of inferring the class label of an *ongoing* (i.e., partially executed, incomplete) action [37]. Thus, our input consists of trimmed video recordings, from which time series of 3D skeletal data can be extracted using a number of methods (e.g., [30])[3] . The observed part of the action depends on the observation ratio which is in the range $(0, 100\%]$. An observation ratio of 100% signifies a fully observed action; in that case, action prediction is equivalent to full action classification.

2 Related work

Vision-based prediction is a rising topic in the field of computer vision [28]. From pedestrian trajectory prediction [33] to pose prediction [22] to accident anticipation [5], prediction has become the focus of several investigations [32]. Ryoo et al. [37] were the first to define the problem of vision-based action prediction as "an inference of unfinished activities given temporally incomplete videos". Action prediction methods can be on trimmed or untrimmed videos [20].

Action prediction in trimmed videos: Action prediction on trimmed videos focuses on recognizing the label of a video given incomplete observations at each point in time [12, 4]. Wang et al. [41] created a teacher-student network where the teacher part of the network recognizes the action while the student part takes as input a partial video and predicts the action. The method in [3] uses a 3D convolutional neural network to extract spatio-temporal features and perform short-term action prediction by using multiple binary classifiers. In [19] the action label is predicted by distinguishing video pairs of different classes that are difficult to discriminate. In [9] everyday actions are predicted by using motion trajectory prediction and taking into account the objects and their affordances.

Action prediction in untrimmed videos: This is performed on non-trimmed videos and its goal is to recognize early (a) all the action labels that are present in a video and (b) their anticipated duration [16, 26, 25]. Since action boundaries are not known, Liu et al.[20] used a scale selection network to select the right window size with the use of convolution layers. The method presented in [11] predicted the future actions with two ways: recursively using an RNN network and in one-pass by using a CNN network. Recursive prediction of actions can accumulate errors. For this reason, Ke et al. [17] predicted actions in an one-shot fashion. The method presented in [13] recognizes the actions of the ongoing video along with their duration by inferring information about the verbs and the objects that are present in the scene through aligning video segments.

Video alignment: Video alignment is gaining increasing popularity in the confrontation of different tasks. The work in [14] uses a variant of DTW with a

[3] In our work the terms "video recordings" and "skeletal data" are used interchangeably.

smoothMin approximation as loss in their network to achieve 3D pose reconstruction and fine-grained audio/visual retrieval. The work in [15] solves the problems of action phase classification, action phase progression, and fine-grained frame retrieval by using temporal alignment and regularization loss. As their temporal alignment loss they employ Soft Dynamic Time Warping (Soft DTW). DTW has been used in the past in relevant contexts. The work in [42] uses DTW for intention prediction based on eye gaze and fixation time-series for autonomous vehicles. In [2] action prediction is based on time-series of features extracted from flow with the use of CNN networks. However, DTW has not been used before for short-term action prediction in the context of human-object interaction scenarios.

Our contribution: We cast the problem of short-term action prediction as a problem of time series alignment and comparison. Video recordings of actions are represented as multidimensional time series through feature extraction mechanisms. The employed features encode the temporal behavior of both the action-related humans and objects. Given a time series representation of an incomplete action, we seek for its best alignment and match with a set of time series corresponding to prototype executions of a variety of actions. The predicted action label is the label of the best matching prototype action execution. For the alignment task, we investigate three DTW variants [38], called OpenEnd-DTW [39], Soft Dynamic Time Warping [8] and Global Alignment Kernel [7]. Our approach is evaluated quantitatively on the skeletal data of three standard benchmark datasets, namely MSR Daily Activities [40], CAD-120 [18] and MHAD [27]. Our results reveal (a) the impact of the used DTW variant and (b) the importance of the fusion of human- and object-centered action representations in the accuracy of action prediction. Moreover, we demonstrate that the proposed approach achieves significantly higher action prediction accuracy compared to competitive methods.

3 DTW-BASED ACTION PREDICTION

We assume that a video recording of an action (full or incomplete) can be represented as an N-dimensional time series through some appropriate feature extraction mechanism (see Sec. 3.1).

Let Q be such a time series representation of an incomplete action. We are interested in inferring the unknown action label $L(Q)$ of Q. We also consider a set of C time series P^i, $1 \leq i \leq C$, corresponding to known, prototype action executions with labels $L(P_i)$. Our approach compares Q with each of the time series P_i through Dynamic Time Warping-based alignment (DTW). Let $DTW(X, Y)$ denote the DTW alignment cost of time series X and Y. Then, action prediction can be formulated as:

$$L(Q) = L\left(arg\ min_{1 \leq i \leq C}\left(DTW(Q, P^i)\right)\right). \qquad (1)$$

Essentially, the proposed method predicts that the action label of Q is that of the prototype action P^i which can be aligned with Q at a minimum DTW-based alignment cost.

3.1 Feature Extraction

Given a video, we represent each of its frames as a multidimensional vector of action-related features. Depending on the employed scenario, such features encode the human body pose, the class and the pose of the involved object, or both. Section 4 presents different sets of extracted features for the standard benchmark datasets employed in this work.

3.2 DTW-based Time Series Alignment

Dynamic Time Warping (DTW): Let X and Y be two time series with $X = (x_1, \dots x_l) \in \mathbb{R}^{n \times l}$ and $Y = (y_1, \dots, y_m) \in \mathbb{R}^{n \times m}$. We define the distance matrix $D(X, Y) = [d(x_i, y_i)]_{ij} \in \mathbb{R}^{l \times m}$, where $d(x, y)$ is the Euclidean distance between x and y. We also define $\mathit{\Pi}$, the set of all path-based alignments of X and Y, connecting the upper-left to the lower-right of the matrix D. Finally, let $\pi \in \mathit{\Pi}$ be one of all those alignments. The inner product $\langle \pi, D(X, Y) \rangle$ yields the alignment score associated with π.

DTW [38] is a dynamic programming algorithm that estimates the minimum-cost alignment of two time series. On the basis of the above notation, this is $DTW(X, Y) = min_{\pi \in \mathit{\Pi}} D(X, Y)$. For the partial alignment, a variant of the original DTW [38] is employed, called open-end DTW [39]. The alignments can end at any point at the last column of the D matrix. The alignment score is normalized by the number of diagonal steps of the calculated optimal alignment path. Open-end DTW is defined as:

$$DTW_{oe}(X, Y) = min_{j=1,\dots,m} DTW(X, Y_j). \tag{2}$$

Soft DTW: Soft-DTW [8] builds upon the original and popular dynamic time warping (DTW) measure and considers a generalized soft minimum operator applied to the distribution of all costs spanned by all possible alignments between two time series of variable size. Given the following generalized minimum operator, subject to a smoothing parameter $\gamma \geq 0$,

$$\min \gamma(\pi_1, \dots, \pi_k) = \begin{cases} \min_{i \leq k} \pi_i, & \gamma = 0, \\ -\gamma \log \sum_{i=1}^{k} e^{\pi_i / \gamma} & \gamma > 0, \end{cases} \tag{3}$$

the soft-DTW score is defined as:

$$SDTW_\gamma(X, Y) = min^\gamma \{\langle \pi, D(X, Y) \rangle, \pi \in \mathit{\Pi}\}. \tag{4}$$

The original DTW score is obtained by setting $\gamma = 0$.

Global Alignment Kernel: Global Alignment Kernel (GAK) [7] measures the similarity between two multidimensional time series X, Y. On top of $D(X, Y)$, the GAK is computed as:

$$GAK(X,Y) = \sum_{\pi \in \mathit{\Pi}} \exp\left(\frac{-\langle \pi, D(X,Y) \rangle}{\gamma}\right). \tag{5}$$

In comparison to DTW, in order to find the alignment score of two time-series, instead of using the operators (min, +) on the $\langle \pi, D(x,y) \rangle$ GAK uses the (+,X) operators. According to [7], the GAK considers the full spectrum of the $\langle \pi, D(X,Y) \rangle, \pi \in \mathit{\Pi}$, while the DTW distance considers only the minimum score.

3.3 Early Fusion of Human and Object Representations

The aforementioned variants of DTW operate on the distance matrix $D(X,Y)$ of time series X and Y, which, for notational convenience, will be denoted with D. In the human-object interaction scenario we are considering, this distance matrix is defined as follows. First, we construct a distance matrix D_H which results from the frame-wise comparison of the part of the representation that contains the information regarding the human. We also construct an analogous distance matrix D_O which results from the frame-wise comparison of the part of the representation that contains the information regarding the object with which the human interacts. Then, the distance matrix D on which DTW operates is defined as:

$$D = \alpha_H D_H + \alpha_O D_O. \tag{6}$$

In the above equation, α_H and α_O are weighting factors that may be dataset-dependent, but have been defined experimentally and commonly for all employed datasets. If the two compared actions involve objects of the same class, then $\alpha_H = \alpha_O = 0.5$. If the two compared actions involve objects of different classes, then $\alpha_H = 0.7$ and $\alpha_O = 0.3$. The intuition behind this choice is that the same action can be performed by using different objects (e.g., reach, move, etc). Therefore, actions can still be compared, but with giving emphasis on the part of the representation concerning the humans rather than the objects. If no objects are present in the scene, then $\alpha_H = 1$ and $\alpha_O = 0$. Finally, in the case that one of the actions involve an object and the other does not, $\alpha_H = 3$ and $\alpha_O = 0$. Essentially, the two actions are again compared on the basis of the human performance, but the mismatch on the presence of objects is penalized by a large α_H value.

In the observed scene, several objects may be present. From those, we consider the one that is closest to and/or manipulated by the actor. One limitation of this choice is that we cannot take into account actions involving more than one object. However, ongoing research beyond the scope of this paper indicates that the extension of our approach towards handling more that one manipulated objects is feasible. Another limitation and future extension of our work lies in the need to know the start frame of an action. The generalization of our approach towards handling unsegmented actions is another topic of ongoing research.

4 EXPERIMENTS

Datasets: The proposed framework is evaluated on 3 standard datasets with different characteristics.

MHAD Dataset [27]: Contains 11 actions (jumping in place, jumping jacks, bending, punching, waving one hand, waving two hands, clapping, throwing a ball, sit down and stand up, sit down, stand up) performed by 12 subjects. The majority of the actions do not involve objects (with the exception of the action "throwing a ball"). The database provides motion capture data containing the 3D positions of 43 LED markers, which have been processed to obtain 3D skeletal data of 30 joints. The standard evaluation split is used as in [27].
Features: The MHAD [27] dataset contains the 3D positions of skeletal joints. Based on these 3D positions, we build a human body representation as proposed in [35] and also used in [29, 21]. Specifically, a human pose is represented as a 30 + 30 + 4 = 64D vector. The first 30 dimensions encode angles of selected body parts with respect to a body-centered coordinate system. The next 30 dimensions encode the same angles in a camera-centered coordinate system. The representation is augmented with the 4 angles between the fore- and the back-arms as well as the angles between the upper- and lower legs.

MSR Daily Activity 3D Dataset [40]: Consists of 16 actions (drinking, eating, reading a book, speaking on cellphone, writing on paper, using a laptop, using a vacuum cleaner, cheering up, sitting still, tossing paper, playing a game, lie down on the sofa, walking, playing the guitar, standing up and sitting down) performed by 10 subjects. Every subject performs each action twice, once sitting on a sofa and once standing. We followed the experimental settings of [43, 34].
Features: The dataset contains the 3D skeletal joint positions for all the human joints. We consider only the 9 upper body joints due to the fact that the data for the lower body are quite noisy. The 3D upper skeletal joint positions are calculated to be invariant to the body center. The invariant 3D joint positions are concatenated with the 3D joint angles. The 3D joint angles are represented as a 30D vector. The 30 dimensions encode angles of selected body parts with respect to a body-centered coordinate system [35] but we are taking into account only the angles that correspond to the upper body. For the objects in this dataset, we employed the YoloV4 [6] trained on ImageNet in order to acquire fast and accurate labels and 2D positions of the objects in the scene. We densely annotated the training part of the MSR-Daily dataset and re-trained the YoloV4 [6] on the MSR Daily Activities dataset. The invariant 3D upper skeletal joint positions are 27D and the 3D joint angles that correspond to the upper body are 18D. The positions of the objects are 2D. Thus, each frame of a video is represented as a 27 + 18 + 2 = 47D vector.

CAD-120 Dataset [18]: Contains activities performed by 4 subjects, which can be subdivided into 10 sub-activities. The subjects perform the activities with different objects. Activities are observed from different viewpoints. The sub-activity labels are: reach, move, pour, eat, drink, open, place, close, clean, null. We are experimenting on the sub-activity labels using the standard 4-fold cross

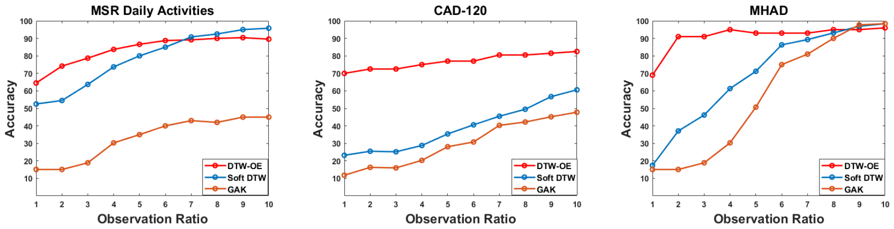

Fig. 1. Comparison of DTW_{oe}, SDTW and GAK on the MSR (left) CAD-120 (middle) and MHAD (right) datasets.

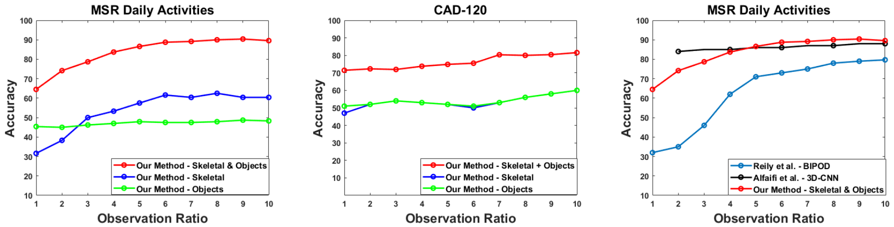

Fig. 2. Action prediction results for different representations. (Left) MSR Daily Activities Dataset, (Middle) CAD-120 Dataset. (Right) Action prediction accuracy of our method in comparison to state of the art methods on the MSR Daily Activities Dataset.

validation as in [18] Features: We used a set of features based on [18]. Specifically, to represent human motion we use the location of each of 8 joints (24D), the distance moved by each joint (8D) and the displacement of each joint (8D). For representing objects we used their 3D centroid location, the distance moved by the object's centroid (1D), the displacement of the object's centroid (1D) and the distance between each joint location and the object centroid (8D). In total, a frame of a sequence is represented as a 53D vector.

Performance metrics: We measure the action prediction accuracy as a function of the observation ratio, i.e., the percentage of the part of the action that has been observed and compared to action prototypes. In our experiments, the observation ratio ranged from 10% to 100% in steps of 10%.

Evaluation of DTW variants: We evaluated the three DTW variants presented in Sec. 3.2 with respect to their action prediction accuracy on all three datasets. During testing, every query sequence is compared to all sequences in the training set. We employed a publicly available [1] implementation of DTW_{oe} and the implementations of the DTW variants that reside in the Tslearn toolkit [36]. The parameter γ was experimentally set equal to 0.1 for the (MHAD, MSR) datasets and to 0.01 for the CAD120 dataset. As it can be observed in Fig. 1, DTW_{oe} outperforms SDTW by a great margin in the MHAD and CAD120 datasets. In turn, SDTW clearly outperforms GAK. This holds true for all three datasets, regardless of whether they involve humans in interaction with objects

(MSR, CAD120) or not (MHAD). Moreover, the superiority of DTW over the rest two variants is dominant especially in lower observation ratios. This shows the potential of the method for accurate and early action prediction.

Evaluation of alternative representations: We evaluated the impact of action representations on action prediction. More specifically, we investigated three different experimental conditions, (a) representations that involve only the joints of the human actor (b) representations that involve only the class and the motion of the involved objects and (c) their early fusion, as presented in Sec. 3.3. Figure 2 (left, middle) shows the results we obtained in the MSR and the CAD120 datasets[4], respectively. As it can be verified, the early fusion of the actor and object representations outperforms any of the individual representations in predictive power, by a vast margin (from a minimum of 10% to a maximum of 40%).

Comparison to the state of the art: Figure 2 (right) presents a comparison of our approach to other competitive methods on the MSR dataset. Specifically, we are comparing to the work of Reily et al. [34] and to that of Alfaifi et al. [3]. As it can be observed, we outperform [34] at all observation ratios and [3] for all observation rations greater than 40%.

To the best of our knowledge, there are no reported quantitative results for action prediction on the CAD120 dataset. We only report action classification results from the very recent method of Mavroudi et al. [24] that achieves an action classification accuracy of 90.4%) which can be compared to the action prediction results of our method in the case of an observation ratio of 100%.

Similarly, there are no reported quantitative results for action prediction on the MHAD dataset. For action classification, the very recent method of Qin et al. [31] achieves an accuracy of 100%, compared to the action classification accuracy of 96% of our method, for an observation ratio of 100%. Interestingly, an action prediction accuracy of more than 90% is achieved by our method, even when a small portion of the activity has been observed (observation ratio of 20%).

5 SUMMARY

We approached the problem of predicting the actions of humans interacting with objects as a problem of aligning fused, frame-based action representations of humans and objects. Specifically, actions are represented as multidimensional time-series. Then, their alignment and the assessment of their similarity is performed with a DTW-based approach. On this task, three DTW variants have been evaluated in three well-known datasets. We also investigated and assessed quantitatively the importance of the fusion of human- and object-based action representations. The obtained results suggest that DTW_{oe} outperforms all tested variants and that the proposed fusion of representations increases considerably

[4] MHAD is not included in this investigation as the vast majority of its actions do not involve human-object interactions.

the predictive capability of our framework, which performs considerably better than recently published competitive action prediction methods.

References

1. https://github.com/statefb/dtwalign
2. Afrasiabi, M., Mansoorizadeh, M., et al.: Dtw-cnn: time series-based human interaction prediction in videos using cnn-extracted features. The Visual Computer (2019)
3. Alfaifi, R., Artoli, A.: Human action prediction with 3d-cnn. SN Computer Science (2020)
4. Arzani, M., Fathy, M., Azirani, A., Adeli, E.: Skeleton-based structured early activity prediction. Multimedia Tools and Applications (2020)
5. Bao, W., Yu, Q., Kong, Y.: Uncertainty-based traffic accident anticipation with spatio-temporal relational learning. In: ACM Int'l Conf. on Multimedia (2020)
6. Bochkovskiy, A., Wang, C., Liao, H.: Yolov4: Optimal speed and accuracy of object detection. arXiv:2004.10934 (2020)
7. Cuturi, M.: Fast global alignment kernels. In: (ICML-11) (2011)
8. Cuturi, M., Blondel, M.: Soft-dtw: a differentiable loss function for time-series. arXiv:1703.01541 (2017)
9. Dutta, V., Zielinska, T.: Predicting human actions taking into account object affordances. Journal of Intelligent & Robotic Systems (2019)
10. Dutta, V., Zielińska, T.: An adversarial explainable artificial intelligence (xai) based approach for action forecasting. Journal of Automation, Mobile Robotics and Intelligent Systems (2021)
11. Farha, A., Richard, A., Gall, J.: When will you do what?-anticipating temporal occurrences of activities. In: IEEE CVPR (2018)
12. Gammulle, H., Denman, S., Sridharan, S., Fookes, C.: Predicting the future: A jointly learnt model for action anticipation. In: IEEE ICCV (2019)
13. Ghoddoosian, R., Sayed, S., Athitsos, V.: Action duration prediction for segment-level alignment of weakly-labeled videos. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2053–2062 (2021)
14. Hadji, I., Derpanis, K.G., Jepson, A.D.: Representation learning via global temporal alignment and cycle-consistency. arXiv preprint arXiv:2105.05217 (2021)
15. Haresh, S., Kumar, S., Coskun, H., Syed, S.N., Konin, A., Zia, M.Z., Tran, Q.H.: Learning by aligning videos in time. arXiv preprint arXiv:2103.17260 (2021)
16. Ke, Q., Bennamoun, M., Rahmani, H., An, S., Sohel, F., Boussaid, F.: Learning latent global network for skeleton-based action prediction. IEEE Trans. on Image Processing (2019)
17. Ke, Q., Fritz, M., Schiele, B.: Time-conditioned action anticipation in one shot. In: IEEE CVPR (2019)
18. Koppula, H., Gupta, R., Saxena, A.: Learning human activities and object affordances from rgb-d videos. The International Journal of Robotics Research (2013)
19. Li, T., Liu, J., Zhang, W., Duan, L.: Hard-net: hardness-aware discrimination network for 3d early activity prediction. In: ECCV (2020)
20. Liu, J., Shahroudy, A., Wang, G., Duan, L., Kot, A.: Skeleton-based online action prediction using scale selection network. IEEE PAMI (2019)
21. Manousaki, V., Papoutsakis, K., Argyros, A.: Evaluating method design options for action classification based on bags of visual words. In: VISAPP (2018)

22. Mao, W., Liu, M., Salzmann, M.: History repeats itself: Human motion prediction via motion attention. In: ECCV (2020)
23. Mavrogiannis, A., Chandra, R., Manocha, D.: B-gap: Behavior-guided action prediction for autonomous navigation. arXiv:2011.03748 (2020)
24. Mavroudi, E., Haro, B., Vidal, R.: Representation learning on visual-symbolic graphs for video understanding. In: ECCV (2020)
25. Miech, A., Laptev, I., Sivic, J., Wang, H., Torresani, L., Tran, D.: Leveraging the present to anticipate the future in videos. In: IEEE CVPR Workshops (2019)
26. Ng, Y., Basura, F.: Forecasting future action sequences with attention: A new approach to weakly supervised action forecasting. IEEE Trans. on Image Processing (2020)
27. Ofli, F., Chaudhry, R., Kurillo, G., Vidal, R., Bajcsy, R.: Berkeley mhad: A comprehensive multimodal human action database. In: IEEE Workshop on Applications of Computer Vision (WACV) (2013)
28. Oprea, S., Martinez-Gonzalez, P., Garcia-Garcia, A., Castro-Vargas, J.A., Orts-Escolano, S., Garcia, J., Argyros, A.: A review on deep learning techniques for video prediction. IEEE PAMI (2020)
29. Papoutsakis, K., Panagiotakis, C., Argyros, A.: Temporal action co-segmentation in 3d motion capture data and videos. In: CVPR (2017)
30. Qammaz, A., Argyros, A.: Occlusion-tolerant and personalized 3d human pose estimation in rgb images. In: 2020 ICPR. IEEE (2021)
31. Qin, Y., Mo, L., Li, C., Luo, J.: Skeleton-based action recognition by part-aware graph convolutional networks. The visual computer (2020)
32. Rasouli, A.: Deep learning for vision-based prediction: A survey. arXiv:2007.00095 (2020)
33. Rasouli, A., Yau, T., Rohani, M., Luo, J.: Multi-modal hybrid architecture for pedestrian action prediction. arXiv:2012.00514 (2020)
34. Reily, B., Han, F., Parker, L., Zhang, H.: Skeleton-based bio-inspired human activity prediction for real-time human–robot interaction. Autonomous Robots (2018)
35. Rius, I., Gonzàlez, J., Varona, J., Roca, F.: Action-specific motion prior for efficient bayesian 3d human body tracking. Pattern Recognition (2009)
36. R.Tavenard, Faouzi, J., Vandewiele, G., Divo, F., Androz, G., Holtz, C., Payne, M., Yurchak, R., Rußwurm, M., Kolar, K., Woods, E.: Tslearn, a machine learning toolkit for time series data. Journal of Machine Learning Research (2020)
37. Ryoo, M.: Human activity prediction: Early recognition of ongoing activities from streaming videos. In: IEEE ICCV (2011)
38. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. IEEE transactions on acoustics, speech, and signal processing (1978)
39. Tormene, P., Giorgino, T., Quaglini, S., Stefanelli, M.: Matching incomplete time series with dynamic time warping: an algorithm and an application to post-stroke rehabilitation. Artificial intelligence in medicine (2009)
40. Wang, J., Liu, Z., Wu, Y., Yuan, J.: Mining actionlet ensemble for action recognition with depth cameras. In: IEEE CVPR (2012)
41. Wang, X., Hu, J., Lai, J., Zhang, J., Zheng, W.: Progressive teacher-student learning for early action prediction. In: IEEE CVPR (2019)
42. Wu, M., Louw, T., Lahijanian, M., Ruan, W., Huang, X., Merat, N., Kwiatkowska, M.: Gaze-based intention anticipation over driving manoeuvres in semi-autonomous vehicles (2020)
43. Xia, L., Aggarwal, J.: Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In: IEEE CVPR (2013)